# Parametric Methods

- Introduction
- Maximum Likelihood Estimation
- Evaluating an Estimator
- Maximum A Posteriori Estimation
- The Bayes' Estimator
- Parametric Classification
- Regression
- Model Selection



# Introduction (1)

- In Bayesian classifier
  - If $P(C_i)$, $p(\mathbf{x}|C_i)$ are unknown
    - Need to estimate from available training data
- Estimation of $p(\mathbf{x}|C_i)$
  - Need sufficient number of training samples $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$
  - 1) Parametric model
    - Parameterizing the densities by an unknown parameter vector $\boldsymbol{\theta}_i$

    $\Rightarrow$ $p(\mathbf{x}|C_i) \equiv p(\mathbf{x}|C_i; \boldsymbol{\theta}_i)$ $\equiv \mathrm{p}(\mathrm{x};\boldsymbol{\theta})$

    $\Rightarrow$ Problem of parameter estimation
  - 2) Nonparametric estimation
    - Without assuming the form of the underlying densities
    - Estimating $p(\mathbf{x}|C_i)$ or $p(C_i|\mathbf{x})$ directly

# Introduction (2)

- Parametric model of $p(\mathbf{x}|C_i) \equiv p(\mathbf{x}|C_i;\boldsymbol{\theta}_i)$

$\Rightarrow$ Problem of parameter estimation $\boldsymbol{\theta}$

- Maximum-likelihood estimation (MLE)
  - The parameters $\boldsymbol{\theta}$ are viewed as fixed quantities but unknown
  - By maximizing the probability of obtaining the samples observed
    - $\widehat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta})$
- Maximum A Posteriori (MAP) estimation
  - The parameters $\boldsymbol{\theta}$ are viewed as random variables having some known prior distribution $p(\boldsymbol{\theta}_i)$
    - $\widehat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{argmax}\, p(\boldsymbol{\theta}|X) = \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$

- This chapter deals with the univariate case, i.e., $\mathbf{x} = [x]$



# Maximum-Likelihood Estimation (1)

- Maximum likelihood parameter estimation
  - Given the training data set
    - $\{X_1, X_2, \ldots, X_K\}$
  - Assumptions
    - The samples in each $X_i$ are i.i.d. following some $p(x|C_i)$
    - $p(x|C_i)$ has a parametric form $p(x|C_i;\boldsymbol{\theta}_i)$
    - Data from one class do not affect the estimation of the others
      - Solving the estimation problem *for each class independently*
  - ML estimation
    - Given a set of training samples $X = \{x_1, \ldots, x_N\}$ drawn independently from the pdf $p(x) \equiv p(x|\boldsymbol{\theta})$
    - To estimate the unknown parameter vector $\boldsymbol{\theta}$
      - By choosing the one that most likely caused the observed data to occur

# Maximum-Likelihood Estimation (2)

- ML estimation of $\boldsymbol{\theta} = [\theta_1, \dots, \theta_r]^T$
  - The likelihood of $\boldsymbol{\theta}$ w.r.t. the set $X$
    - $p(X|\boldsymbol{\theta}) = p(x_1, x_2, \dots, x_N|\boldsymbol{\theta})$
      $= \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})$
      - Once $X$ is given, $p(X|\boldsymbol{\theta})$ is a function of θ alone
  - ML estimate of θ
    - $\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{argmax} \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})$
    - Log-likelihood function
      - $L(\boldsymbol{\theta}) \equiv ln\, p(X|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(x_i|\boldsymbol{\theta})$
    - $\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{argmax}\, L(\boldsymbol{\theta})$
      - Let $\nabla_{\boldsymbol{\theta}} L \equiv \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$
        » The solution could be a true global maximum, a local maximum (or minimum) or an inflection point of $L(\theta)$

For small sample size, mode is in general a poorer estimate than mean or median

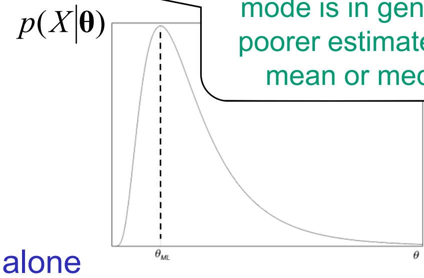



Fig. 2.14 [Theodoridis 09]

The gradient operator

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$



# Maximum-Likelihood Estimation (3)

- Bernoulli distribution
  - In a Bernoulli distribution
    - There are only 2 possible outcomes: $\{1,0\}$ with probabilities $p, (1-p)$
    - $p(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \\ 0, & \text{otherwise} \end{cases}$
    - $p(x) = p^x (1-p)^{1-x}, \quad x \in \{0,1\}$
      - $p\ (0 < p < 1)$ is the only parameter
  - $E(X) = \sum_x x p(x) = 1 \times p + 0 \times (1-p) = p$
  - $Var(X) = \sum_x \left(x - E(X)\right)^2 p(x) = E(X^2) - [E(X)]^2$
    $= p - p^2 = p(1-p)$

# Maximum-Likelihood Estimation (4)

- Bernoulli distribution (cont.)
  - Given an iid sample $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \{0,1\}$
  - To calculate the estimator $\hat{p}$
    - $L(p) = \sum_{i=1}^{N} \ln p(x_i|p) = \sum_{i=1}^{N} (x_i \ln p + (1 - x_i) \ln(1 - p))$
      $= \ln p \sum_{i=1}^{N} x_i + \ln(1 - p) \, (N - \sum_{k=1}^{N} x_i)$
    - $\frac{\partial L(p)}{\partial p} = 0$
      - $\hat{p}_{ML} = \frac{\sum_{i=1}^{N} x_i}{N}$
      - The estimate is the ratio of (#occurrences of the event) to (#experiments)



# Maximum-Likelihood Estimation (5)

- Multinomial distribution
  - The outcome is one of $K$ disjoint states with probabilities $p_1, \dots, p_K$
    - $p_1 + \cdots + p_K = \sum_{k=1}^{K} p_k = 1$
  - Let $\mathbf{x} = [x_1, \dots, x_K]^T$ be the indicator vector
    - $x_k \in \{0,1\}, \sum_{k=1}^{K} x_k = 1$
      - $x_k = 1$ if the outcome is state $k$ and $x_k = 0$ otherwise
    - $p(\mathbf{x}) = p(x_1, x_2, \dots, x_K) = p_1^{x_1} p_2^{x_2} \dots p_K^{x_K} = \prod_{k=1}^{K} p_k^{x_k}$
  - Given an iid sample $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
    - $L(p_1, \dots, p_K) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i | p_1, \dots, p_K) = \sum_{i=1}^{N} \sum_{k=1}^{K} x_{i,k} \ln p_k$
      $= \sum_{k=1}^{K} \ln p_k \sum_{i=1}^{N} x_{i,k}$
    - By the Lagrange multipliers $\mathcal{L} = L(p_1, \dots, p_K) + \lambda(1 - \sum_{k=1}^{K} p_k)$
    - ML estimates $\hat{p}_k = \frac{\sum_{i=1}^{N} x_{i,k}}{N}$
      - (#occurrences of the outcomes of state $k$) / (#experiments)

# Maximum-Likelihood Estimation (6)

- Uniform distribution
  - Assume $X = \{x_1, x_2, \dots, x_N\}$ are drawn from a uniform distribution in the interval $(0, \theta)$
    - $p(x|\theta) \sim U(0, \theta) = \begin{cases} \frac{1}{\theta}, & 0 < x \le \theta \\ 0, & otherwise \end{cases}$
      - where $\theta$ is unknown
  - The likelihood function is
    - $p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \frac{1}{\theta^N}, \quad 0 < x_i \le \theta, i = 1, \dots N$

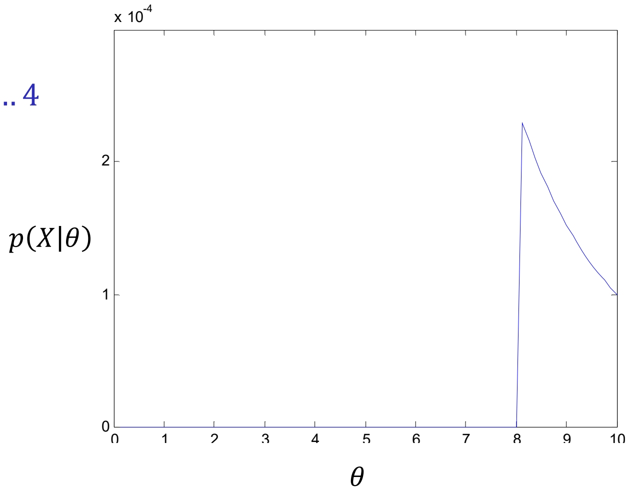
      $= \frac{1}{\theta^N}, \quad 0 < \max(x_1, \dots, x_N) \le \theta$
  - The ML estimate for $\theta$ is
    - $\hat{\theta}_{ML} = \max(x_1, \dots, x_N)$



# Maximum-Likelihood Estimation (7)

- Uniform distribution (cont.)
  - Example (p.98, [Duda, 01])
    - Given an iid sample $X = \{4,7,2,8\}$ drawn from a uniform distribution
      - $p(x|\theta) \sim U(0, \theta) = \begin{cases} \frac{1}{\theta}, & 0 < x \le \theta \le 10 \\ 0, & otherwise \end{cases}$
    - Then
      - $p(X|\theta) = \frac{1}{\theta^4}, \quad 0 < x_i \le \theta, i = 1, \dots 4$

        $= \frac{1}{\theta^4}, \quad 8 \le \theta \le 10$
      - $\hat{\theta}_{ML} = 8$



# Maximum-Likelihood Estimation (8)

- The univariate Gaussian Case 1: unknown $\mu$
  - Suppose the samples are drawn from $N(\mu, \sigma^2)$
    - $L(\theta) = L(\mu) = \sum_{i=1}^{N} \ln p(x_i|\mu) = \sum_{i=1}^{N} \left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}\right\}$
      $= -\frac{N}{2}\ln(2\pi) - N\ln\sigma - \sum_{i=1}^{N} \left\{\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$
  - Thus
    - $\frac{\partial L(\mu)}{\partial \mu} = \sum_{i=1}^{N} \left\{\frac{x_i-\mu}{\sigma^2}\right\} = 0$
    - $\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$
    - The sample mean is the ML optimal for Gaussians
      (NOT necessarily ML optimal for non-Gaussians)

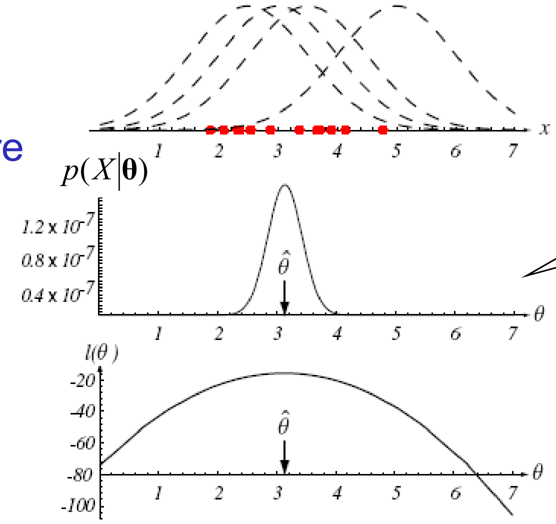


# Maximum-Likelihood Estimation (9)

- The univariate Gaussian Case 1: unknown $\mu$
  - $\hat{\mu}_{ML}$ is a unbiased estimate of the mean
    - $E\{\hat{\mu}_{ML}\} = E\left\{\frac{1}{N}\sum_{i=1}^{N} x_i\right\} = \frac{1}{N}\sum_{i=1}^{N} E\{x_i\} = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$
    - $bias(\hat{\mu}_{ML}) = E\{\hat{\mu}_{ML}\} - \mu = 0$
  - Variance of the estimate $\hat{\mu}_{ML}$
    - $Var[\hat{\mu}_{ML}] = E\{(\hat{\mu}_{ML} - \mu)^2\}$
      $= \frac{1}{N^2}\sum_{i=1}^{N} E\{(x_i-\mu)^2\} + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i} E\{\{(x_i-\mu)\{(x_j-\mu)\}$
      $= \frac{\sigma^2}{N}$

# Maximum-Likelihood Estimation (10)

- Example [Duda, 01]
  - Suppose the samples are drawn from $N(\mu, \sigma^2)$
    - Unknown μ




a function of θ

Fig. 3.1 [Duda 01]

FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of $\theta$ whereas the conditional density $p(x|\theta)$ is shown as a function of $x$. Furthermore, as a function of $\theta$, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



# Maximum-Likelihood Estimation (11)

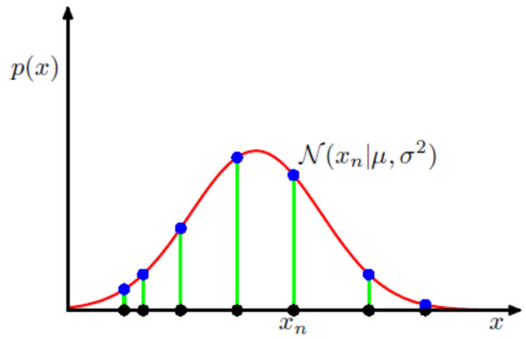
- The univariate Gaussian Case 2: unknown $\mu$ and $\sigma^2$
  - $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$
  - $L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \left\{ -\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x_i-\theta_1)^2}{2\theta_2} \right\}$

    $= -\frac{N}{2}\ln(2\pi\theta_2) - \sum_{i=1}^{N} \left\{ \frac{(x_i-\theta_1)^2}{2\theta_2} \right\}$
  - Let $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\sum_{i=1}^{N}(x_i-\theta_1)}{\theta_2} \\ -\frac{N}{2\theta_2} + \frac{\sum_{i=1}^{N}(x_i-\theta_1)^2}{2\theta_2^2} \end{bmatrix} = \mathbf{0}$
    - $\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$
    - $\hat{\sigma}^2{}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{ML})^2$

# Maximum-Likelihood Estimation (12)

- Example
  - Fig. 1.14 [Bishop 06]

**Figure 1.14** Illustration of the likelihood function for a Gaussian distribution, shown by the red curve. Here the black points denote a data set of values $\{x_n\}$, and the likelihood function given by (1.53) corresponds to the product of the blue values. Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.




- data $\{x_1, x_2, \dots, x_N\}$ : black points
- $p(x_i|\boldsymbol{\theta})$: blue points
- Likelihood function: product of the blue values

$$p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

- ML: adjusting the mean and variance so as to maximize the product



# Maximum-Likelihood Estimation (13)

- The univariate Gaussian Case 2: unknown $\mu$ and $\sigma^2$
  - $\hat{\mu}_{ML}$ is a unbiased estimate of the mean
    - $E\{\hat{\mu}_{ML}\} = E\left\{\frac{1}{N}\sum_{i=1}^{N} x_i\right\} = \frac{1}{N}\sum_{i=1}^{N} E\{x_i\} = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$
    - $bias(\hat{\mu}_{ML}) = E\{\hat{\mu}_{ML}\} - \mu = 0$
  - $\hat{\sigma}^2{}_{ML}$ is a biased estimate of the variance for finite $N$
    - $E\{\hat{\sigma}^2{}_{ML}\} = E\left\{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{ML})^2\right\} = \frac{1}{N}\sum_{i=1}^{N} E\{(x_i - \hat{\mu}_{ML})^2\}$

$$= \frac{1}{N}\sum_{i=1}^{N} E\{((x_i - \mu) - (\hat{\mu}_{ML} - \mu))^2\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\{(x_i - \mu)^2 - 2(x_i - \mu)(\hat{\mu}_{ML} - \mu) + (\hat{\mu}_{ML} - \mu)^2\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(\sigma^2 - \frac{2\sigma^2}{N} + \frac{\sigma^2}{N}\right)$$

$$= \frac{N-1}{N}\sigma^2 \neq \sigma^2$$

# Maximum-Likelihood Estimation (14)

- The univariate Gaussian Case 2: unknown $\mu$ and $\sigma^2$
  - ${\hat{\sigma}^2}_{ML}$ is a biased estimate of the variance for finite $N$
    - $E\{{\hat{\sigma}^2}_{ML}\} = \frac{N-1}{N}\sigma^2 \neq \sigma^2$
    - However, for large $N$,
      - $E\{{\hat{\sigma}^2}_{ML}\} \approx \sigma^2$
  - In Matlab
    - $var(X)$ returns ${\hat{\sigma}^2}_{N-1}$
    - ${\hat{\sigma}^2}_{N-1} = \frac{N}{N-1}{\hat{\sigma}^2}_{ML} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}_{ML})^2$
    - ${\hat{\sigma}^2}_{N-1}$ is a unbiased estimate of the variance
      - $E\{{\hat{\sigma}^2}_{N-1}\} = E\left\{\frac{N}{N-1}{\hat{\sigma}^2}_{ML}\right\} = \frac{N}{N-1}\frac{N-1}{N}\sigma^2 = \sigma^2$
    - $var(X, 1)$ returns ${\hat{\sigma}^2}_{ML}$



# Maximum-Likelihood Estimation (15)

- Example
  - Fig. 1.15 [Bishop 06]
    - Data are generated from the distribution in green curve
    - Three data sets
      - Each consists of 2 blue points
      - ML results are shown in red curves
    - Averaged across the 3 data sets
      - The mean is correct
        - » $E\{\hat{\mu}_{ML}\} = \mu$
      - The variance is under-estimated because it is measured relative to the sample mean but not the true mean
        - » $E\{{\hat{\sigma}^2}_{ML}\} = \frac{N-1}{N}\sigma^2$

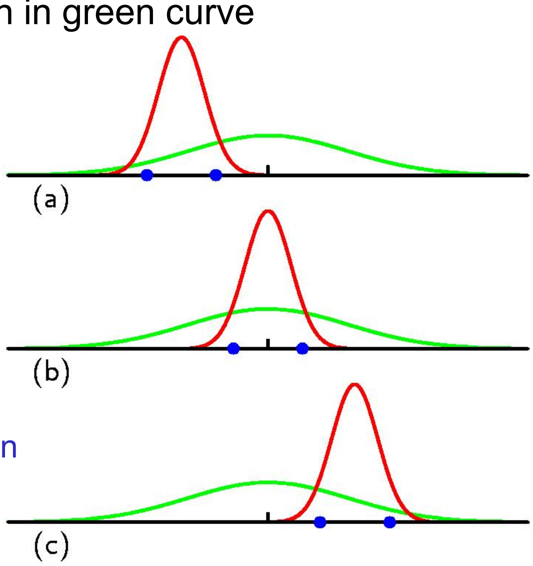



Fig. 1.15 [Bishop 06]

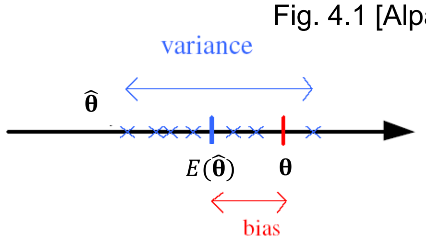
# Evaluating an Estimator (1)

- The estimator $\hat{\boldsymbol{\theta}}$
  - Let $X$ be a sample from a population specified up to a parameter $\boldsymbol{\theta}$
  - To evaluate the quality of an estimator $\hat{\boldsymbol{\theta}}$
    - The estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X)$ is a random variable
      - Because it depends on the sample $X$
    - The evaluation should be averaged over all possible $X \sim P(X|\boldsymbol{\theta})$
- The bias of an estimator $\hat{\boldsymbol{\theta}}$
  - $\text{bias}(\hat{\boldsymbol{\theta}}) = E_{P(X|\boldsymbol{\theta})}\{\hat{\boldsymbol{\theta}}(X)\} - \boldsymbol{\theta} = E_{P(X|\boldsymbol{\theta})}\{\hat{\boldsymbol{\theta}}(X) - \boldsymbol{\theta}\}$
- Unbiased estimator
  - If $\text{bias}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$
    - $E_{P(X|\boldsymbol{\theta})}\{\hat{\boldsymbol{\theta}}(X)\} = \boldsymbol{\theta}$
    - The sampling distribution is centered on the true parameter



# Evaluating an Estimator (2)

Fig. 4.1 [Alpaydin, 2014]




- The variance of an estimator $\hat{\boldsymbol{\theta}}$
  - $var(\hat{\boldsymbol{\theta}}) = E\left[\left(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})\right)^2\right]$
    - Being unbiased is not enough
    - An estimate may be unbiased, but the resulting estimates may exhibit large variations around the mean
- Consistent estimator
  - If the estimator eventually recovers the true parameters as the sample size $N$ goes to infinity
    - $\lim_{N\to\infty} P\left(\left|\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}\right| \le \varepsilon\right) = 1, \ \forall \varepsilon > 0$
      - $\hat{\boldsymbol{\theta}}(X) \to \boldsymbol{\theta}$ *as* $N \to \infty$
  - Or, for large $N$, the variance of the estimate tends to zero
    - $\lim_{N\to\infty} E\left[\left(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})\right)^2\right] = 0$

# Evaluating an Estimator (3)

- The mean square error of the estimator $\hat{\boldsymbol{\theta}}$
  - $r(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\right] = E\left[(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}}) + E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta})^2\right]$

    $= E\left[\left(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})\right)^2\right] + E\left[(E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta})^2\right] + 2E[(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}}))(E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta})]$

    $= E\left[\left(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})\right)^2\right] + \left(E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right)^2 + 2\left(E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right)E[\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})]$

    $= E\left[\left(\hat{\boldsymbol{\theta}} - E(\hat{\boldsymbol{\theta}})\right)^2\right] + \left(E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right)^2$

    $= var(\hat{\boldsymbol{\theta}}) + \mathrm{bias}(\hat{\boldsymbol{\theta}})^2$
  - Even if the estimator is unbiased, it can still result in a large MSE due to a large variance term

The bias-variance tradeoff

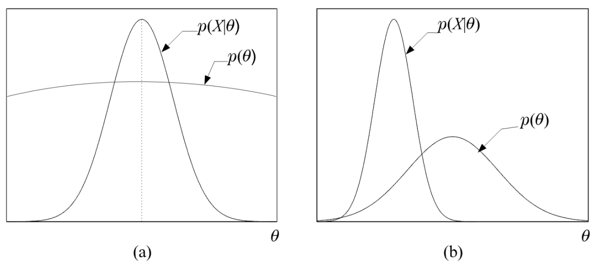


# MAP Estimation (1)

- In ML estimation of $\boldsymbol{\theta}$
  - $\boldsymbol{\theta}$ is considered as an unknown nonrandom parameter vector
- In MAP estimation
  - $\boldsymbol{\theta}$ is considered as a random vector described by a known pdf $p(\boldsymbol{\theta})$
  - Given a set of training samples $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$
    - Finding the maximum of posterior
    - $\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{argmax}\, p(\boldsymbol{\theta}|X)$

      $= \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$
    - Let $\nabla_{\boldsymbol{\theta}}(p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})) = 0$
  - If $p(\boldsymbol{\theta})$ is uniform or flat enough
    - $\hat{\boldsymbol{\theta}}_{MAP} \cong \hat{\boldsymbol{\theta}}_{ML}$

Fig. 2.7 [Theodoridis 09]




$\hat{\boldsymbol{\theta}}_{MAP} \cong \hat{\boldsymbol{\theta}}_{ML}$ $\qquad$ $\hat{\boldsymbol{\theta}}_{MAP} \neq \hat{\boldsymbol{\theta}}_{ML}$

# MAP Estimation (2)

- Example 2.5 [Theodoridis 09]
  - The Gaussian Case: unknown $\mu$
    - Suppose the samples are drawn from $N(\mu, \sigma^2)$
    - Suppose the unknown $\mu$ is known to be normally distributed
      - $p(\theta) = p(\mu) \sim N(\mu_0, \sigma_0^2) = \frac{1}{(2\pi\sigma_0^2)^{\frac{1}{2}}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$
    - The posterior density
      - $p(\theta|X) = p(\mu|X) = \frac{p(X|\mu)p(\mu)}{\int p(X|\mu)p(\mu)d\mu} = \alpha\left[\prod_{i=1}^{N} p(x_i|\mu)\right]p(\mu) = \cdots = N(\mu_N, \sigma_N^2)$
      - where
        - » $\mu_N = \frac{N\sigma_0^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2} = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$
        - » $\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$

$\alpha = P(X)$ is a normalization factor that depends on $X$ but is independent of $\mu$



# MAP Estimation (3)

- Example 2.5 (cont.)
  - The Gaussian Case: unknown $\mu$
    - MAP estimation
      - $\because p(\mu|X) = N(\mu_N, \sigma_N^2)$
      - Let $\frac{\partial p(\mu|X)}{\partial \mu} = 0$
        - » $\Rightarrow \hat{\mu}_{MAP} = \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$
    - Or from $\frac{\partial \ln(p(X|\mu)p(\mu))}{\partial \mu} = 0$
      - $\Rightarrow \frac{\partial \ln(\prod_{i=1}^{N} p(x_i|\mu)p(\mu))}{\partial \mu} = \frac{\partial \ln(p(\mu))}{\partial \mu} + \sum_{i=1}^{N} \frac{\partial}{\partial \mu} \ln p(x_i|\mu) = 0$
      - $\Rightarrow -\frac{1}{\sigma_0^2}(\mu - \mu_0) + \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu) = 0$
      - $\Rightarrow \hat{\mu}_{MAP} = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$

The posterior mean is a linear combination of ML mean and the prior mean $\mu_0$

# MAP Estimation (4)

- Example 2.5 (cont.)
  - The Gaussian Case: unknown $\mu$
    - $\hat{\mu}_{MAP} = \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$
    - If the prior is weak, i.e., $\sigma_0^2 \gg \sigma^2$ or $N \to \infty$
      - $\hat{\mu}_{MAP} \approx \hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$
    - If the prior is strong, i.e., $\sigma_0^2 = 0$ or $N \to 0$
      - $\hat{\mu}_{MAP} \approx \mu_0$

The posterior $p(\mu|X) = N(\mu_N, \sigma_N^2)$ with different number of training samples

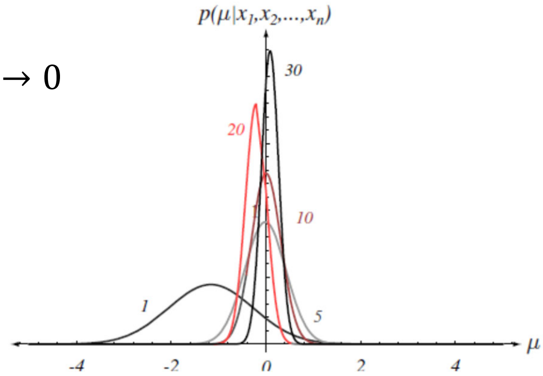



Fig. 3.2 [Duda 01]



# MAP Estimation (5)

- Comparison: ML & MAP
  - The univariate Gaussian Case: unknown $\mu$
  - ML
    - $\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$
    - $bias(\hat{\mu}_{ML}) = 0$
    - $var(\hat{\mu}_{ML}) = \frac{\sigma^2}{N}$
  - MAP
    - $\hat{\mu}_{MAP} = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 = w\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + (1-w)\mu_0$
      - $0 \le w \le 1$
    - $bias(\hat{\mu}_{MAP}) = w\mu + (1-w)\mu_0 - \mu = (1-w)(\mu_0 - \mu)$
    - $var(\hat{\mu}_{MAP}) = w^2\frac{\sigma^2}{N}$
    - Although the MAP estimate is biased, it has lower variance

# MAP Estimation (6)

- Example (Gaussian case, Fig. 4.12 [Murphy 2012])
  - Given a noisy observation $x = 3$
    - Likelihood $p(x|\mu) = N(x|\mu, \sigma^2) = N(x|\mu, 1)$
    - Prior $p(\mu) = N(0, \sigma_0^2) = N(0,1)$ and $p(\mu) = N(0,5)$
    - Posterior $p(\mu|x) = N\left(\frac{3}{2}, \frac{1}{2}\right)$ and $p(\mu|x) = N\left(\frac{15}{6}, \frac{5}{6}\right)$

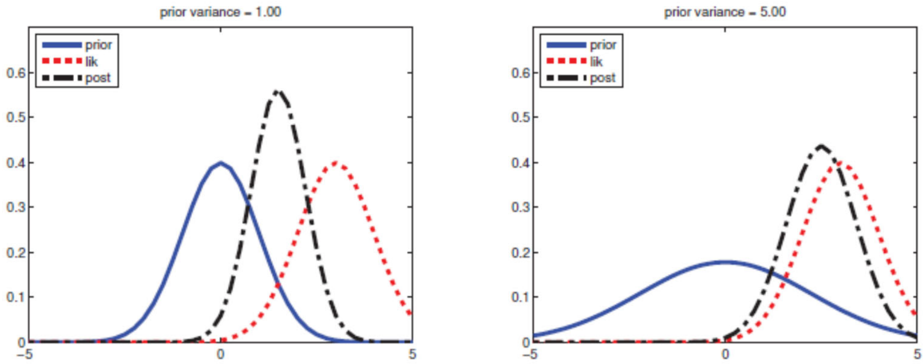



Fig. 4.12 [Murphy 2012]

**Figure 4.12** Inference about $x$ given a noisy observation $y = 3$. (a) Strong prior $\mathcal{N}(0, 1)$. The posterior mean is "shrunk" towards the prior mean, which is 0. (a) Weak prior $\mathcal{N}(0, 5)$. The posterior mean is similar to the MLE. Figure generated by `gaussInferParamsMean1d`.



# MAP Estimation (7)

- In ML & MAP
  - Estimation of the unknown θ
    - $\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta})$
    - $\hat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{argmax}\, p(\boldsymbol{\theta}|X) = \underset{\boldsymbol{\theta}}{argmax}\, p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})$
  - Drawbacks
    - No measure of uncertainty
      - How much one can trust an estimate
    - Can result in overfitting
    - The mode is an untypical point

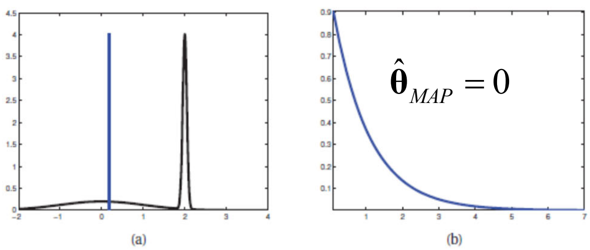



Fig. 5.1 [Murphy 2012]

**Figure 5.1** (a) A bimodal distribution in which the mode is very untypical of the distribution. The thin blue vertical line is the mean, which is arguably a better summary of the distribution, since it is near the majority of the probability mass. Figure generated by `bimodalDemo`. (b) A skewed distribution in which the mode is quite different from the mean. Figure generated by `gammaPlotDemo`.

# The Bayes' Estimator (1)

- Bayes' estimator (or Bayesian Inference)
  - Given $X = \{x_1, x_2, \dots, x_N\}$ and $p(\boldsymbol{\theta})$
  - Estimation of *the posterior predicative* pdf $p(x|X)$
    - Instead of taking a single estimate of $\boldsymbol{\theta}$
    - Taking the average of $p(x|\boldsymbol{\theta})$, weighted by $p(\boldsymbol{\theta}|X)$, over all possible $\boldsymbol{\theta}$
    - $p(x|X) = \int p(x,\boldsymbol{\theta}|X)d\boldsymbol{\theta} = \int p(x|\boldsymbol{\theta},X)p(\boldsymbol{\theta}|X)d\boldsymbol{\theta} = \int p(x|\boldsymbol{\theta})p(\boldsymbol{\theta}|X)d\boldsymbol{\theta}$

The distribution of $x$ is known completely once we know $\boldsymbol{\theta}$

Use of the full $p(\boldsymbol{\theta}|X)$ distribution

    - Where the posterior density
      » $p(\boldsymbol{\theta}|X) = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(X)} = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$
    - By the independent assumption
      » $p(X|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})$



# The Bayes' Estimator (2)

- $p(x|X) = \int p(x|\boldsymbol{\theta})p(\boldsymbol{\theta}|X)d\boldsymbol{\theta}$
  - $p(\boldsymbol{\theta})$ is called prior density of $\boldsymbol{\theta}$ (prior to the measurements)
  - $p(\boldsymbol{\theta}|X)$ is called posterior density of $\boldsymbol{\theta}$ (after the measurements)
    - $p(\boldsymbol{\theta}|X) = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(X)} = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$
  - The likelihood function $p(X|\boldsymbol{\theta})$ is considered as a function of $\boldsymbol{\theta}$ (not a density)
    - $p(X|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})$
  - If $p(\boldsymbol{\theta}|X)$ peaks very sharply about some $\widehat{\boldsymbol{\theta}}$
    - Then $p(x|X) \cong p(x|\widehat{\boldsymbol{\theta}})$

# The Bayes' Estimator (3)

- Gaussian case – univariate with unknown $\mu$
  - Assume $p(x|\theta) = p(x|\mu) \sim N(\mu, \sigma^2)$
    - With the known prior density
      - $p(\theta) = p(\mu) \sim N(\mu_0, \sigma_0^2)$
  - The posterior density
    - After observing an iid sample $X = \{x_1, x_2, \ldots, x_N\}$
    - $p(\theta|X) = p(\mu|X) = \frac{p(X|\mu)p(\mu)}{\int p(X|\mu)p(\mu)d\mu} = \alpha\left[\prod_{i=1}^{N} p(x_i|\mu)\right]p(\mu)$

$\alpha = P(X)$ is a normalization factor that depends on $X$ but is independent of $\mu$

$$= \alpha\left[\prod_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] \frac{1}{(2\pi\sigma_0^2)^{\frac{1}{2}}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i + \frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right]$$



# The Bayes' Estimator (4)

- Gaussian case – univariate with unknown $\mu$ (cont.)
  - The posterior density $p(\mu|X)$ is again a normal density
    - An exponential function of a quadratic function of $\mu$
      - $p(\mu|X) = \frac{1}{(2\pi\sigma_N^2)^{\frac{1}{2}}} \exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right) \sim N(\mu_N, \sigma_N^2)$
        - » $\mu_N = \frac{N\sigma_0^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}$
        - » $\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2}$

$\mu_N$ represents our best guess for $\mu$ after observing $N$ samples

$\sigma_0 = 0 \Longrightarrow \mu_N \to \mu_0$

$\sigma_0 \gg \sigma \Longrightarrow \mu_N \to \frac{1}{N}\sum_{i=1}^{N} x_i$

$\sigma_N^2$ measures our uncertainty about the true value of $\mu$

- As $N \to \infty$
  - $\mu_N \to \hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - $\sigma_N^2 \to \frac{\sigma^2}{N}$
  - $p(\mu|X)$ becomes more sharply peaked around the same mean

Each additional observation decreases our uncertainty about the true value of $\mu$

# The Bayes' Estimator (5)

- Gaussian case – univariate with unknown $\mu$ (cont.)
  - The desired class-conditional density $p(x|X)$
    - $p(x|X) = \int p(x|\mu)p(\mu|X)d\mu$

$$= \int \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \frac{1}{(2\pi\sigma_N^2)^{\frac{1}{2}}} \exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right) d\mu$$

$$= \frac{1}{(2\pi\sigma^2\sigma_N^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{(x-\mu_N)^2}{\sigma^2+\sigma_N^2}\right) \int \exp\left(-\frac{1}{2}\frac{\sigma^2+\sigma_N^2}{\sigma^2\sigma_N^2}\left(\mu - \frac{\sigma^2\mu_N+\sigma_N^2 x}{\sigma^2+\sigma_N^2}\right)^2\right) d\mu$$

$$\sim N(\mu_N, \sigma^2 + \sigma_N^2)$$

  - That is, given $p(x|\mu) \sim N(\mu, \sigma^2)$
    - $p(x|X) \sim N(\mu_N, \sigma^2 + \sigma_N^2)$

The increased variance results from our lack of exact knowledge of $\mu$

- $p(x|X)$ is $p(x|C_i, X)$ in the classifier design
- Bayesian classifier: $\max_{C_i}\{P(C_i|x, X_i)\} = \max_{C_i}\{p(x|C_i, X_i)P(C_i)\}$



# The Bayes' Estimator (6)

- Example 2.6 (p. 40, [Theodoridis 09])
  - Bayesian learning of the unknown mean $\mu$
    - The data were generated with $p(x|\mu) \sim N(\mu, \sigma^2) = N(2,4)$
    - The prior adopted is $p(\mu) \sim N(\mu_0, \sigma_0^2) = N(0,8)$
  - The posterior
    - $p(\mu|X) \sim N(\mu_N, \sigma_N^2)$
    - $N \to \infty$
      - $\mu_N \to \hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$

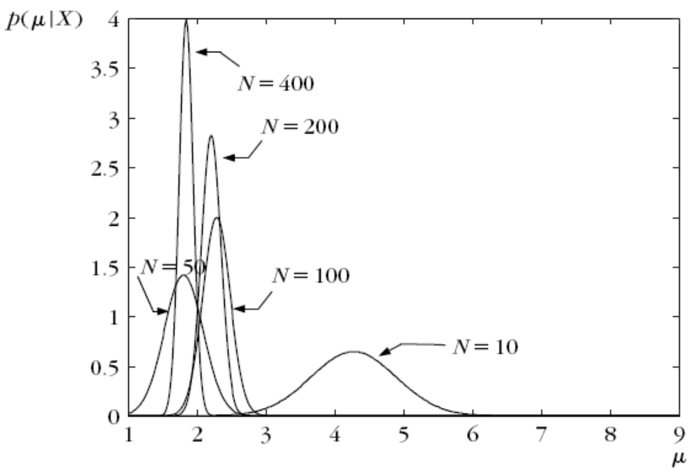



Fig. 2.16 [Theodoridis 09]

# The Bayes' Estimator (7)

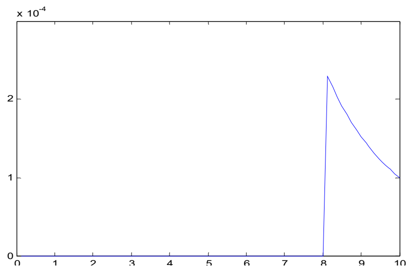
- Relation to MAP/ML solution
  - If $p(\boldsymbol{\theta}|X)$ is sharply peaked at $\widehat{\boldsymbol{\theta}}_{MAP}$
  - Then
    - $p(x|X) \approx p(x|\widehat{\boldsymbol{\theta}}_{MAP})$
  - e.g.,
    - If $p(X|\boldsymbol{\theta})$ is concentrated around $\widehat{\boldsymbol{\theta}}_{ML}$ and $p(\boldsymbol{\theta})$ is flat enough
      - $\widehat{\boldsymbol{\theta}}_{MAP} \cong \widehat{\boldsymbol{\theta}}_{ML}$
      - $p(x|X) \approx p(x|\widehat{\boldsymbol{\theta}}_{MAP}) \approx p(x|\widehat{\boldsymbol{\theta}}_{ML})$

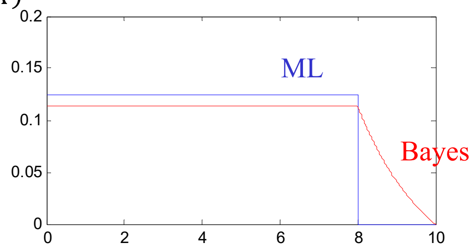


# The Bayes' Estimator (8)

- Example (p.98, [Duda, 01])
  - Given $X = \{4,7,2,8\}$ drawn from a uniform distribution
    - $p(x|\theta) \sim U(0,\theta) = \begin{cases} \frac{1}{\theta}, & 0 < x \le \theta \le 10 \\ 0, & otherwise \end{cases}$
      - $p(X|\theta) = \frac{1}{\theta^4}, \quad 0 < x_i \le \theta, i = 1, \dots 4$

        $= \frac{1}{\theta^4}, \quad 8 \le \theta \le 10$
      - $\hat{\theta}_{ML} = 8$
    - Assume $p(\theta) \sim U(0,10)$
      - $p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \propto \begin{cases} \frac{1}{\theta^4}, & 8 \le \theta \le 10 \\ 0, & otherwise \end{cases}$
      - $p(x|X) = \begin{cases} \int_8^{10} \frac{1}{\theta} p(\theta|X)\, d\theta, & 0 \le x \le 8 \\ \int_x^{10} \frac{1}{\theta} p(\theta|X)\, d\theta, & x > 8 \end{cases}$



# Parametric Classification (1)

- In Chap 3
  - $g_i(x) = ln\, p(x|C_i) + \ln P(C_i)$
  - Assuming $g_i(x) \sim N(\mu_i, \sigma_i^2)$
    - $g_i(x) = -\frac{(x-\mu_i)^2}{2\sigma_i^2} - \frac{1}{2}\ln(2\pi) - \ln \sigma_i + \ln P(C_i)$
  - We estimate the unknown parameters for each class separately
    - $\hat{\mu}_{i,ML} = \frac{1}{N}\sum_{j=1}^{N} x_j$
    - $\hat{\sigma}_i{}^2{}_{ML} = \frac{1}{N}\sum_{j=1}^{N}\left(x_j - \hat{\mu}_{i,ML}\right)^2$
    - $\hat{P}(C_i) = \frac{N_i}{N_1 + \cdots + N_K}$
  - The discriminant function becomes
    - $g_i(x) = -\frac{(x-\hat{\mu}_i)^2}{2\hat{\sigma}_i{}^2} - \frac{1}{2}\ln(2\pi) - \ln \hat{\sigma}_i + \ln \hat{P}(C_i)$



# Parametric Classification (2)

- Example (Ch3, case 1)
  - $\sigma_i^2 = \sigma_j^2$
  - $P(C_i) = P(C_j)$
  - $g_i(x) = -(x - \hat{\mu}_i)^2$

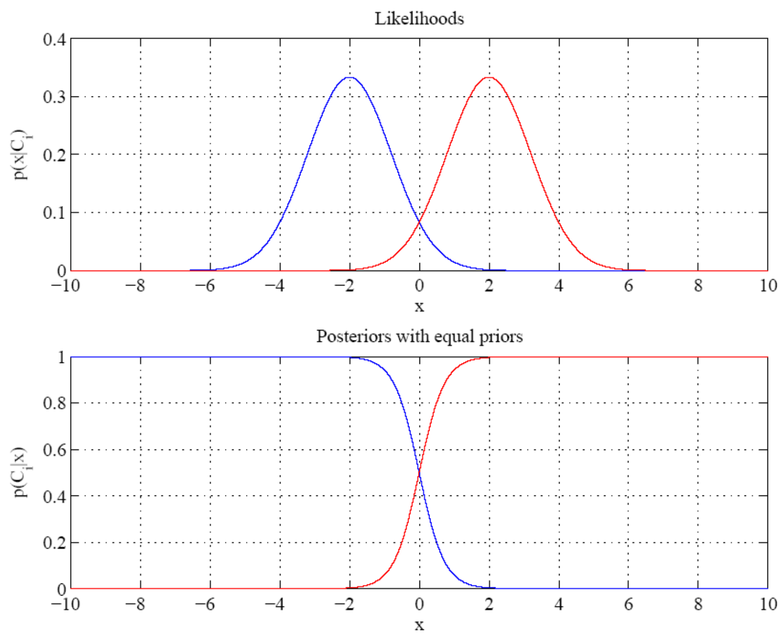



Fig. 4.2 [Alpaydin]

# Parametric Classification (3)

- Example (Ch3, case 4)
  - $\sigma_i^2 \neq \sigma_j^2$
  - $P(C_i) = P(C_j)$
  - $g_i(x) = \frac{(x-\mu_i)^2}{2\sigma_i^2} - \ln \sigma_i$

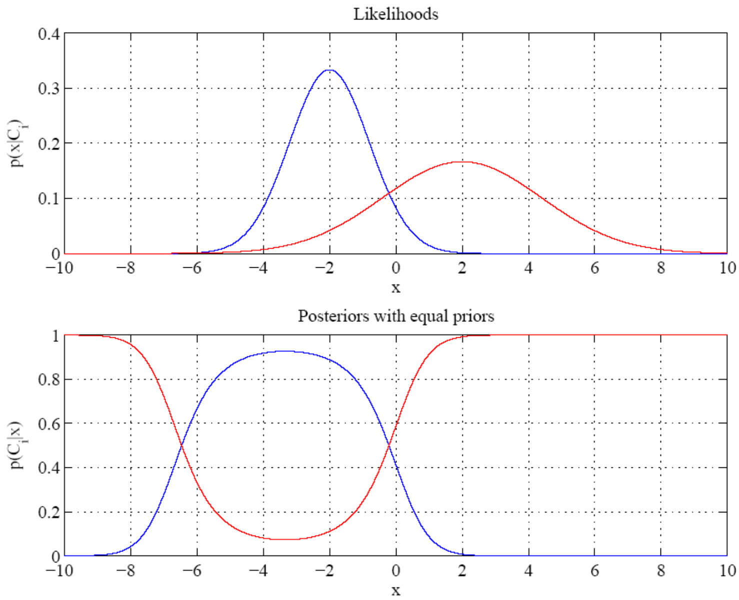



Fig. 4.3 [Alpaydin]



# Regression (1)

- Regression
  - $X = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$
    - $y = f(x) + \varepsilon$, $\varepsilon$ is random noise and is assumed to be $\varepsilon \sim N(0, \sigma^2)$
  - To approximate the unknown $f(x)$ by the estimator $g(x|\mathbf{\theta})$
    - $p(y|x, \mathbf{\theta}) \sim N(y|g(x|\mathbf{\theta}), \sigma^2)$
  - The log-likelihood
    - $L(\boldsymbol{\theta}) \equiv ln\, p(X|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(y_i|x_i, \boldsymbol{\theta})$

$$= \sum_{i=1}^{N} \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{\left(y_i - g(x_i|\boldsymbol{\theta})\right)^2}{2\sigma^2}\right)\right)$$

$$= -\frac{N}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - g(x_i|\boldsymbol{\theta})\right)^2$$

  - Maximizing $L(\boldsymbol{\theta})$ = minimizing the sum of squared error
    - $E(\boldsymbol{\theta}|X) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - g(x_i|\boldsymbol{\theta})\right)^2$

# Regression (2)

- Linear regression
  - Assuming that $g(x|\boldsymbol{\theta})$ is linear
    - $g(x|w_1, w_0) = w_1 x + w_0$
  - The sum of squared error
    - $E(w_1, w_0|X) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - g(x_i|\boldsymbol{\theta})\right)^2 = \frac{1}{2}\sum_{i=1}^{N}(y_i - w_1 x_i - w_0)^2$
  - Let
    - $\frac{\partial}{\partial w_0}E = 0 \Rightarrow \sum_{i=1}^{N} y_i = N w_0 + w_1 \sum_{i=1}^{N} x_i$
    - $\frac{\partial}{\partial w_1}E = 0 \Rightarrow \sum_{i=1}^{N} y_i x_i = w_0 \sum_{i=1}^{N} x_i + w_1 \sum_{i=1}^{N} x_i^2$
    - $\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} y_i x_i \end{bmatrix}$

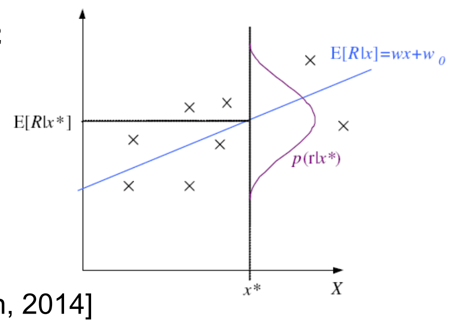



Fig. 4.4 [Alpaydin, 2014]



# Regression (3)

- Polynomial regression
  - Assuming that $g(x|\boldsymbol{\theta})$ is a polynomial in $x$ of order $k$
  - $g(x|w_k, \ldots, w_0) = w_k x^k + \cdots + w_2 x^2 + w_1 x + w_0$
  - $E(\boldsymbol{w}|X) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$
    - $\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \ldots & x_1^k \\ \vdots & \ddots & & \vdots \\ 1 & x_N & \ldots & x_N^k \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix}$
    - $\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0$
      - $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$ (normal equation)
      - $\widehat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

# Regression (4)

- Square error
  - $E(\boldsymbol{\theta}|X) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - g(x_i|\boldsymbol{\theta})\right)^2$
- Relative square error
  - $E_{RSE}(\boldsymbol{\theta}|X) = \frac{\sum_{i=1}^{N}\left(y_i - g(x_i|\boldsymbol{\theta})\right)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
    - If $E_{RSE} \to 0$
      - We have better fit
    - If $E_{RSE} \to 1$
      - The prediction is as good as predicting by the average
      - Using a model based on input $x$ does not work better than using the average



# Regression (5)

- The expected square error at $x$
  - $E\left(\left(y - g(x)\right)^2 \middle| x\right) = \underbrace{E((y - E(y|x))^2|x)}_{\text{Noise}} + \underbrace{(E(y|x) - g(x))^2}_{\text{Squared error}}$
    - $E((y - E(y|x))^2|x)$: variance of $y$ given $x$
      - The variance of noise added, not depends on $g(x)$
    - $(E(y|x) - g(x))^2$: how much $g(x)$ deviates from $E(y|x)$
      - This term depends on the estimator and the training set $X$
- The expected value over samples $X$
  - $E_X\left(\left(E(y|x) - g(x)\right)^2 \middle| x\right)$
    $= \underbrace{E_X[(g(x) - E_X(g(x))^2]}_{\text{Variance}} + \underbrace{(E(y|x) - E_X(g(x)))^2}_{\text{Bias}}$
    - Bias measures how much $g(x)$ is wrong
    - Variance measures how much $g(x)$ fluctuate around $E_X(g(x))$

# Regression (6)

- Estimating bias and variance
  - From a number of datasets $X_j$ $(j = 1, \dots, M)$
    - Using each $X_j$ to form an estimator $g_j(\cdot)$
    - $E_X(g(x))$ is estimated by the average over $g_j(\cdot)$
      - $\bar{g}(x) = \frac{1}{M}\sum_{j=1}^{M} g_j(x)$
  - $bias^2(g) = \frac{1}{N}\sum_{i=1}^{N}[\bar{g}(x_i) - f(x_i)]^2$
  - $var(g) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[g_j(x_i) - \bar{g}(x_i)\right]^2$
- Two examples
  - A constant fit $g_j(x) = 2, \forall j$
    - No variance & high bias
  - Taking the average $g_j(x) = \frac{1}{N}\sum_{i=1}^{N} y_i$
    - Lower bias & increased variance



# Regression (7)

- Example
  - $f(x) = 2\sin(1.5x)$
  - $y = f(x) + \varepsilon,$
    - $\varepsilon \sim N(0,1)$
  - 5 datasets
    - $X_1, \dots, X_5$
    - $M = 5$
    - $N = 20$
  - The dotted line
    - $\bar{g}(x)$

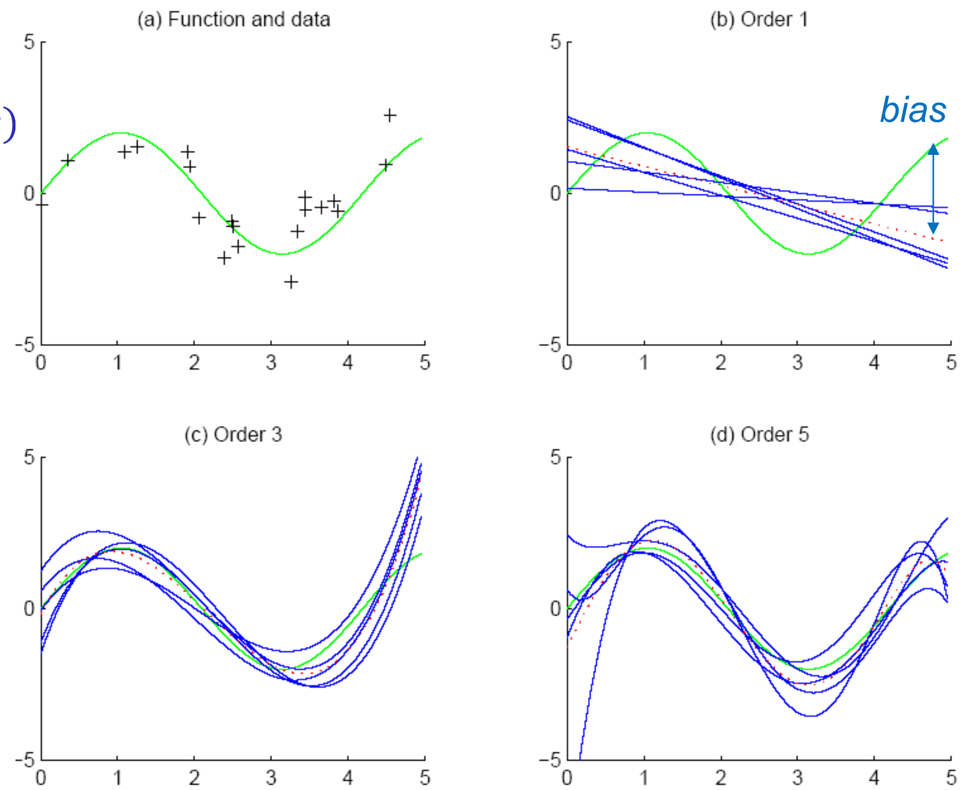



Fig. 4.5 [Alpaydin]

# Regression (8)

- Bias-variance dilemma
  - As we increase complexity
    - Bias decreases
      - A better fit to data
    - Variance increases
      - Fit varies more with data
- Example
  - $M = 100$
  - Order 1
    - The smallest variance
  - Order 5
    - The smallest bias
  - Order 3
    - The minimum error

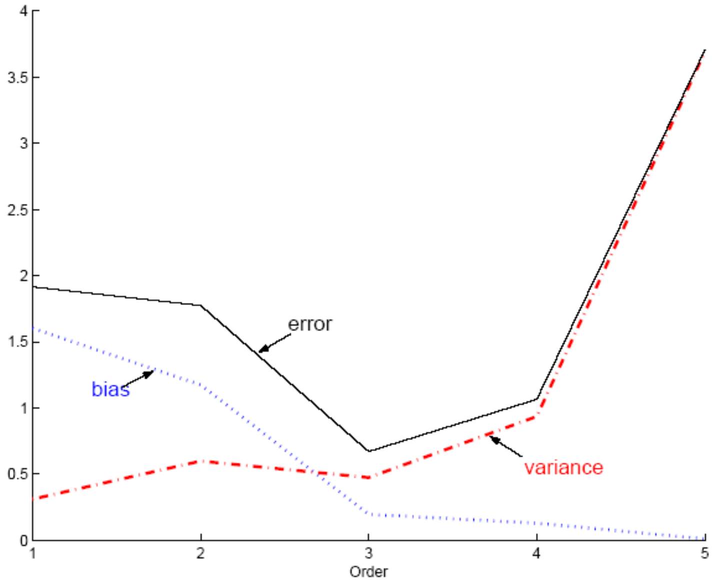
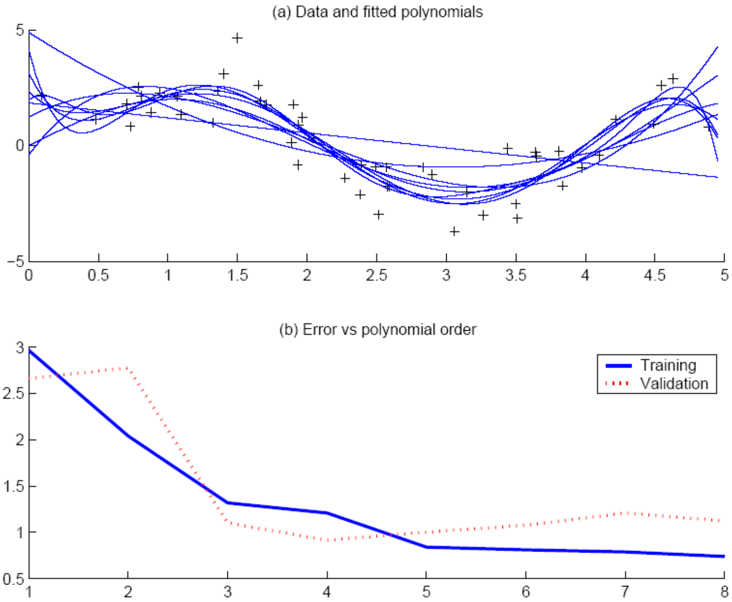



Fig. 4.6 [Alpaydin]



# Model Selection (1)

- Cross validation
  - In practice, we cannot calculate the bias and variance for a model
  - But we can calculate the validation error as an estimate




50 training instances
50 validation instances
Polynomial order: 1, 2, …,8

The elbow is at 3

Fig. 4.7 [Alpaydin]

# Model Selection (2)

- Regularization
  - The augmented error function
    - $E = \text{error on data} + \lambda \cdot \text{model complexity}$
      - The 2nd term penalizes complex models with large variances
      - If λ is too large, only very simple models are allowed and bias ↑
        - » λ is determined using cross-validation
  - Example (L2 regularization)
    - In the regression model $g(x|\boldsymbol{w})$
    - $E = \sum_{i=1}^{N}\left(y_i - g(x_i|\boldsymbol{w})\right)^2 + \lambda \sum_k w_k^2$

Coefficients increase in magnitude as order increases:
1: [-0.0769, 0.0016]
2: [0.1682, -0.6657, 0.0080]
3: [0.4238, -2.5778, 3.4675, -0.0002
4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]

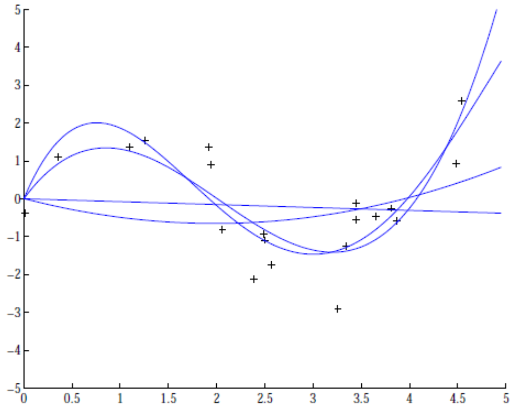



Fig. 4.8 [Alpaydin]